On monotone 'metrics' of the classical channel space:non-asymptotic theory

Keiji Matsumoto
National Institute of Informatics, Tokyo, Japan

February 1, 2010

Abstract

The aim of the manuscript is to characterize monotone 'metric' in the space of Markov map. Here, 'metric' means the square of the norm defined on the tangent space, and not necessarily induced from an inner product (this property hereafter will be called inner-product-assumption), different from usual metric used in differential geometry.

As for metrics in So far, there have been plenty of literatures on the metric in the space of probability distributions and quantum states. Among them, Cencov proved the monotone metric in probability distribution space is unique up to constant multiple, and identical to Fisher information metric. Petz characterized all the monotone metrics in the quantum state space using operator mean. As for channels, however, only a little had been known.

In this paper, we impose monotonicity by concatenation of channels before and after the given channel families, and invariance by tensoring identity channels. (Notably, we do *not* use the inner-product-assumption.) To obtain this result, 'resource conversion' technique, which is widely used in quantum information, is used. We consider distillation from and formation to a family of channels. Under these axioms, we identify the largest and the smallest 'metrics'. Interestingly, they are *not* induced from any inner product, i.e., not a metric. Indeed, one can prove that *any* 'metric' satisfying our axioms can *not* be a metric.

This result has some impact on the axiomatic study of the monotone metric in the space of classical and quantum states, since both conventional theory relies on the inner-product-assumption. Also, we compute the lower and the upper bound for some concrete examples.

1 Introduction

The aim of the manuscript is to characterize monotone 'metric' in the space of Markov map. Here, 'metric' means the square of the norm defined on the tangent space, and not necessarily induced from an inner product, different from usual metric used in differential geometry.

So far, there have been plenty of literatures on the metric in the space of probability distributions and quantum states. Cencov, sometime in 1970s, proved the monotone metric in probability distribution space is unique up to constant multiple, and identical to Fisher information metric [4]. He also discussed invariant connections in the same space. Amari and others independently worked on the same objects, especially from differential geometrical view points, and applied to number of problems in mathematical statistics, learning theory, time series analysis, dynamical systems, control theory, and so on[1][2]. Quantum mechanical states are discussed in literatures such as [2][3][5][5][6]. Among them Petz [6] characterized all the monotone metrics in the quantum state space using operator mean.

As for channels, however, only a little had been known. To my knowledge, there had been no study about axiomatic characterization of distance measures in the classical or quantum channel space.

In this paper, we impose monotonicity by concatenation of channels before and after the given channel families, and invariance by tensoring identity channels. (Notably, we do *not* use the inner-product-assumption.) To obtain this result, 'resource conversion' technique, which is widely used in quantum information, is used. We consider distillation from and formation to a family of channels.

Under these axioms, we identify the largest and the smallest 'metric'. Interestingly, they are *not* induced from any inner product, i.e., not a metric. Indeed, one can prove that *any* 'metric' satisfying our axioms can *not* be a metric.

In author's opinion, the axiom in this manuscript is reasonable and minimal, and it is essential that being metric in narrow sense is not required. Hence, this result has some impact on the axiomatic study of the monotone metric in the space of classical and quantum states, since both Cencov [4] and Petz [6] relies on the inner-product-assumption. Since classical and quantum states can be viewed as channels with the constant output, it is preferable to dispense with the inner-product-assumption. This point will be discussed in a separate manuscript.

2 Notations and conventions

- $\otimes_{\rm in}$ ($\otimes_{\rm out}$) :the totality of the input (output) alphabet

- $\mathcal{P}_{\rm in}$ ($\mathcal{P}_{\rm out}$) : the totality of the probability distributions over $\otimes_{\rm in}$ ($\otimes_{\rm out}$). In this paper, the existence of density with respect to an underlying measure μ is always assumed. Hence, $\mathcal{P}_{\rm in}$ ($\mathcal{P}_{\rm out}$) is equivalent

to the totality of density functions.

- $\mathcal{C}$: the totality of channels which sends an element of $\mathcal{P}_{\mathrm{in}}$ to an element of $\mathcal{P}_{\mathrm{out}}$
- $\mathcal{P}_k$: totality of probability mass functions supported on $\{1, 2, \cdots, k\}$
- $\mathcal{C}_{k,l}$: totality of the Markov map from $\mathcal{P}_k$ to $\mathcal{P}_l$
- x,y, etc.: an element of $\otimes_{\mathrm{in}}$,$\otimes_{\mathrm{out}}$
- X,Y, etc.: random variable taking values in $\otimes_{\mathrm{in}}$,$\otimes_{\mathrm{out}}$
- A probability distribution p is identified with the Markov map which sends all the input probability distributions to p. (Hence represented by a transition matrix of rank 1.)
- $\mathcal{T}_{\cdot}(\cdot)$: tangent space
- δ etc. : an element of $\mathcal{T}_p(\mathcal{P}_{\mathrm{in}})$ etc.
- Δ etc. : an element of $\mathcal{T}_\Phi(\mathcal{C})$
- An element δ of $\mathcal{T}_p(\mathcal{P}_{\mathrm{in}})$ etc. is identified with an element f of L^1 such that $\int f \mathrm{d}\mu = 0$.
- $g_p(\delta)$: square of a norm in $\mathcal{T}_p(\mathcal{P}_k)$
- $G_\Phi(\Delta)$: square of a norm in $\mathcal{T}_\Phi(\mathcal{C}_{k,l})$
- $J_p(\delta)$: classical Fisher information
- The local data at p: the pair $\{p, \delta\}$.
- The local data at Φ : the pair $\{\Phi, \Delta\}$.
- $\Phi(\cdot|x) \in \mathcal{P}_{\mathrm{out}}$: the distribution of the output alphabet when the input is x
- $\Delta(\cdot|x) \in \mathcal{T}_p(\mathcal{P}_{\mathrm{out}})$ is defined as the infinitesimal increment of above
- $\mathbf{I}$: identity

3 Axioms

(M1) $G_\Phi(\Delta) \geq G_{\Phi\circ\Psi}(\Delta \circ \Psi)$

(M2) $G_\Phi(\Delta) \geq G_{\Psi\circ\Phi}(\Psi \circ \Delta)$

(E) $G_{\Phi\otimes\mathbf{I}}(\Delta \otimes \mathbf{I}) = G_\Phi(\Delta)$

(N) $G_p(\delta) = g_p(\delta)$

4 Programming or simulation of channel families

Suppose we have to fabricate a channel Φ_θ, which is drawn from a family $\{\Phi_\theta\}$, without knowing the value of θ but with a probability distribution q_θ or a channel Ψ_θ, drawn from a family $\{q_\theta\}$ or $\{\Psi_\theta\}$. More specifically, we need a channel Λ with

$$\Phi_\theta = \Lambda \circ (\mathbf{I} \otimes q_\theta), \tag{1}$$

or channels Λ_a and Λ_b with

$$\Phi_\theta = \Lambda_b \circ (\Psi_\theta \otimes \mathbf{I}) \circ \Lambda_a. \tag{2}$$

Here, note that Λ, Λ_a, and Λ_b should not vary with the parameter θ. Note also that the former is a special case of the latter. Also, giving the value of θ with infinite precision corresponds to the case of $q_\theta = \delta(x-\theta)$.

Differentiating the both ends of (1) and (2), and letting $\Phi_\theta = \Phi$, $q_\theta = q$, and $\Psi_\theta = \Psi$, we obtain

$$\Delta = \Lambda \circ (\mathbf{I} \otimes \delta), \tag{3}$$

and

$$\Delta = \Lambda_b \circ (\Delta' \otimes \mathbf{I}) \circ \Lambda_a, \tag{4}$$

where $\Delta \in \mathcal{T}_\Phi(\mathcal{C}_{k,l})$, $\delta \in \mathcal{T}_q(\mathcal{P}_{k'})$, and $\Delta' \in \mathcal{T}_\Psi(\mathcal{C}_{k',l'})$.

In the manuscript, we consider *tangent simulation*, or the operations satisfying (1) (or (2)) and (3) (or (4), resp.), at the point $\Phi_\theta = \Phi$ only. Especially, we are interested in point simulation of the 1-dimensional subfamily. Note that simulation of $\{\Phi, \Delta\}$ is equivalent to the one of the channel family $\{\Phi_{\theta+t} = \Phi + t\Delta\}_t$.

5 Relation between g and G

In this section, we study norms with (M1), (M2), (E), and (N).

Theorem 1 *Suppose (M1) and (N) hold. Then,*

$$G_\Phi(\Delta) \geq G_\Phi^{\min}(\Delta) := \sup_{p\in\mathcal{P}_{\rm in}} g_{\Phi(p)}(\Delta(p)) = \max_{x\in\Omega_{\rm in}} g_{\Phi(\cdot|x)}(\Delta(\cdot|x)).$$

Also, $G_\Phi^{\min}(\Delta)$ satisfies (M1), (M2), (E), and (N).

Proof.

$$G_\Phi(\Delta) = G_\Phi(\Delta) \geq G_{\Phi\circ p}(\Delta \circ p) = g_{\Phi(p)}(\Delta(p)).$$

The last identity is trivial. Obviously, $G_\Phi^{\min}(\Delta)$ satisfies (M1), (M2) and (N). (E) is seen from the right most side expression. ■

Theorem 2 *Suppose (M2), (E) and (N) hold. Then*

$$G_{\Phi}(\Delta) \leq G_{\Phi}^{\max}(\Delta) := \inf_{\Lambda, q, \delta} \{ g_q(\delta) \,;\ \Lambda \circ (\mathbf{I} \otimes q) = \Phi,\, \Lambda \circ (\mathbf{I} \otimes \delta) = \Delta \ \}.$$

Also, $G_{\Phi}^{\max}(\Delta)$ satisfies (M1), (M2), (E), and (N).

Proof.

$$\begin{aligned} g_q(\delta) = G_q(\delta) = G_{\mathbf{I}\otimes q}(\mathbf{I} \otimes \delta) &\geq G_{\Lambda \circ (\mathbf{I}\otimes q)}(\Lambda \circ (\mathbf{I} \otimes \delta)) \\ = G_{\Phi}(\Delta). \end{aligned}$$

So we have the inequality. That $G_{\Phi}^{\max}(\Delta)$ satisfies (M1), (M2), (E), and (N) is trivial. ■

Corollary 3

$$G_{\Phi}^{\max}(\Delta) \geq G_{\Phi}^{\min}(\Delta).$$

Obviously, $G_{\Phi}^{\min}(\Delta)$ and $G_{\Phi}^{\max}(\Delta)$ are not induced from any metric, i.e., they cannot be written as $S(\Delta, \Delta)$, where S is a positive real bilinear form. Indeed, we can show the following theorem:

Theorem 4 *Suppose (M1), (M2), (E) and (N) hold. For any interior point Φ of $\mathcal{C}_{2,2}$, $G_{\Phi}(\Delta)$ cannot written as $S_{\Phi}(\Delta, \Delta)$, where S_{Φ} is a positive real bilinear form.*

Proof. Let Φ be the one which corresponds to the stochastic matrix

$$\begin{bmatrix} 1-t & s \\ t & 1-s \end{bmatrix}.$$

Also, let

$$\Delta_1 := \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \Delta_2 := \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}.$$

Since the family $\{\Phi + \theta\Delta_1\}_{\theta}$ can be simulated by the simulation suggested by the decomposition

$$\Phi + \theta\Delta_1 = (1 - t + \theta)(\Phi + t\Delta_1) + (t - \theta)(\Phi - (1 - t)\Delta_1),$$

(M2) and (E), we have $G_{\Phi}(\Delta_1) \leq g_p(\delta)$, where $p = (1-t, t)$ and $\delta = (1, -1)$. On the other hand, by chosing input as $(1, 0)$, $\{\Phi, \Delta_1\}$ induces $\{p, \delta\}$. Therefore, by (M1), $G_{\Phi}(\Delta_1) \geq g_p(\delta)$ and hence

$$G_{\Phi}(\Delta_1) = g_p(\delta).$$

Similarly, we have

$$G_{\Phi}(\Delta_2) = g_q(\delta'),$$

where $q = (s, 1-s)$ and $\delta' = (1, -1)$. Consider the family $\{\Phi + t(\Delta_1 + a\Delta_2)\}_t$. If $|a| < \min\left\{\frac{1-s}{t}, \frac{s}{t}, \frac{1-s}{1-t}, \frac{s}{1-t}\right\}$, this can be generated by the simulation suggested by

$$\Phi + t(\Delta_1 + a\Delta_2) = (1 - t + \theta)(\Phi + t\Delta_1 + ta\Delta_2) + (t - \theta)(\Phi - (1-t)\Delta_1 - a(1-t)\Delta_2).$$

Therefore, $G_\Phi(\Delta_1 + a\Delta_2) \leq g_p(\delta)$. On the other hand, by chosing input as $(1, 0)$, $\{\Phi, \Delta_1 + a\Delta_2\}$ induces $\{p, \delta\}$. Therefore,

$$G_\Phi(\Delta_1 + a\Delta_2) = g_p(\delta).$$

On the other hand, if $G_\Phi(\Delta) = S_\Phi(\Delta, \Delta)$ with some linear bilinear form S_Φ,

$$\begin{aligned} G_\Phi(\Delta_1 + a\Delta_2) &= S_\Phi(\Delta_1 + a\Delta_2, \Delta_1 + a\Delta_2) \\ &= S_\Phi(\Delta_1, \Delta_1) + a^2 S_\Phi(\Delta_2, \Delta_2) + 2aS_\Phi(\Delta_1, \Delta_2) \\ &= g_p(\delta) + a^2 g_q(\delta') + 2aS(\Delta_1, \Delta_2). \end{aligned}$$

Hence, it should hold that

$$a^2 g_q(\delta') + 2aS_\Phi(\Delta_1, \Delta_2) = 0$$

for any $|a| < \min\left\{\frac{1-s}{t}, \frac{s}{t}, \frac{1-s}{1-t}, \frac{s}{1-t}\right\}$. Hence, $g_q(\delta') = 0$. Since $\delta \neq 0$, this is contradiction. ■

Observe that the argument parallel with the above proof applies also to $\mathcal{C}_{k,l}$ $(k,l \geq 3)$. The following property isuseful in computation of $G^{\max}$.

Proposition 5 *Let $\left\{\Upsilon^{(i)}\right\}_{i=1}^n$ be the extreme points of $\mathcal{C}$. Then*

$$G_\Phi^{\max}(\Delta) = \min_{q,\delta} g_q(\delta)$$

where $q = (q_1, \cdots, q_n)$ is a probability distribution over $\left\{\Upsilon^{(i)}\right\}$ with

$$\Phi = \sum_{i=1}^n q_i \Upsilon^{(i)},$$

and $\delta = (\delta_1, \cdots, \delta_n)$ satisfies $\Delta = \sum_{i=1}^n \delta_i \Upsilon^{(i)}$.

Proof. Consider a simulation suggested by the decomposition

$$\Phi = \int \Psi \mathrm{d}P(\Psi), \quad \Delta = \int \Psi f \mathrm{d}P(\Psi),$$

where P is a probability measure over $\mathcal{C}$ and $\int f \mathrm{d}P(\Psi) = 0$. Here the 'program' is $\{P, f \circ P\}$, where $f \circ P$ is the singed measure defined by $f \circ P(A) = \int_A f \mathrm{d}P(\Psi)$. Letting $\Psi = \sum_{i=1}^n p_{i|\Psi} \Upsilon^{(i)}$, we obtain another simulation corresponding to the decomposition

$$\Phi = \sum_i q_i \Upsilon^{(i)}, \quad \Delta = \sum_i \delta_i \Upsilon^{(i)},$$

where

$$q_i := \int p_{i|\Psi} \mathrm{d}P(\Psi), \, \delta_i := \int p_{i|\Psi} f \mathrm{d}P(\Psi).$$

Here the 'program' is the pair $\{q, \delta\}$. The following Markov map sends the pair $\{P, f \circ P\}$ to the pair $\{q, \delta\}$: upon accepting Ψ, which is generated according to the probability measure P, generate $\Upsilon^{(i)}$ with the probability $p_{i|\Psi}$. Therefore, by monotonicity,

$$g_P(f \circ P) \geq g_q(\delta),$$

which implies the assertion. ■

6 Binary channels $\mathcal{C}_{2,2}$

In this section, we suppose g is the Fisher information metric. $\mathcal{C}_{2,2}$ has four extreme points,

$$\Upsilon^{(1)} := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Upsilon^{(2)} := \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \Upsilon^{(3)} := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \Upsilon^{(4)} := \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

and can be parameterized as

$$\begin{bmatrix} 1-t & s \\ t & 1-s \end{bmatrix}.$$

Hence the space can be viewed as a square. Consider one-parameter subfamily $\{\Phi_\theta\}$ of $\mathcal{C}_{2,2}$, passing through Φ. Let Ψ_A and Ψ_B the intersection of the edge of $\mathcal{C}_{2,2}$ and the tangent line at Φ with the tangent Δ. Obviously, $\{\Phi, \Delta\}$ can be simulated as a probabilistic mixture of Ψ_A and Ψ_B. Hence, defining a and b by $\Delta = a(\Psi_A - \Psi_B)$ and $\Phi = b\Psi_A + (1-b)\Psi_B$,

$$G_\Phi^{\max}(\Delta) \leq \frac{a^2}{b} + \frac{a^2}{1-b}.$$

Suppose Ψ_A and Ψ_B can be discriminated with certainty by observing the output for a properly chosen input. This occurs if and only if one of the following is true:

$$\begin{aligned} [\Psi_A]_{11} &= 1 \;\; \& \quad [\Psi_B]_{01} = 1, \\ [\Psi_A]_{01} &= 1 \;\; \& \quad [\Psi_B]_{11} = 1, \\ [\Psi_A]_{10} &= 1 \;\; \& \quad [\Psi_B]_{00} = 1, \\ [\Psi_A]_{00} &= 1 \;\; \& \quad [\Psi_B]_{10} = 1. \end{aligned}$$

In such cases, one can extract the Fisher information of the binary distribution which is used to mix Ψ_A and Ψ_B. Therefore,

$$G_\Phi^{\min}(\Delta) \geq \frac{a^2}{b} + \frac{a^2}{1-b}.$$

Hence, due to Corollary 3, we have

$$G_{\Phi}(\Delta)=G_{\Phi}^{\min}(\Delta)=G_{\Phi}^{\max}(\Delta)=\frac{a^2}{b}+\frac{a^2}{1-b}.$$

Especially, if $\Phi=\frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$, this is the case for any Δ.

In general, however, the simulation by the mixture of Ψ_A and Ψ_B is not optimal. For example, let

$$\begin{aligned}\Phi &:= a\Upsilon^{(1)}+b\Upsilon^{(2)}+c\Upsilon^{(3)}=(a-t)\,\Upsilon^{(1)}+(b+t)\,\Upsilon^{(2)}+(c-t)\,\Upsilon^{(3)}+t\Upsilon^{(4)}\\ &=\begin{bmatrix}a & c\\ b+c & a+b\end{bmatrix}=\begin{bmatrix}a & c\\ 1-a & 1-c\end{bmatrix},\\ \Delta &:= \begin{bmatrix}-1 & 1\\ 1 & -1\end{bmatrix}=\Upsilon^{(3)}-\Upsilon^{(1)}=(1-s)\,\Upsilon^{(3)}+s\left(\Upsilon^{(2)}+\Upsilon^{(4)}-\Upsilon^{(1)}\right)-\Upsilon^{(1)}\\ &=-(1+s)\,\Upsilon^{(1)}+s\Upsilon^{(2)}+(1-s)\,\Upsilon^{(3)}+s\Upsilon^{(4)},\end{aligned}$$

with

$$a+b+c=1,\;\; 0\le t\le 1,\;\; s\in\mathbb{R}$$

We use Proposition 5.

$$G_{\Phi}^{\max}(\Delta)=\min_{\substack{s\in\mathbb{R}\\ t\in[0,\min\{a,c\}]}}\left[\frac{(1+s)^2}{a-t}+\frac{s^2}{b+t}+\frac{(1-s)^2}{c-t}+\frac{s^2}{t}\right]$$

First, we optimize over s, which achieves minimum at

$$s=\frac{(a-c)\,t\,(t+b)}{-t^2+2act+abc}.$$

Hence,

$$\begin{aligned}G_{\Phi}^{\max}(\Delta)&=\min_{t\in[0,\min\{a,c\}]}\frac{2t+ab+bc}{-t^2+2act+abc}\\ &=\min_{t\in[0,\min\{a,c\}]}\frac{2t+ab+bc}{\left(\left(ac+\sqrt{a^2c^2+abc}\right)-t\right)\left(t-\left(ac-\sqrt{a^2c^2+abc}\right)\right)}\end{aligned}$$

After some computation, one can verify

$$ac+\sqrt{a^2c^2+abc}=ac+\sqrt{a^2c^2+ac\,(1-a-c)}\le\min\{a,c\}.$$

Therefore, the function to be optimized is monotone increasing in the domain. Hence, the minimum is achieved at $t=0$. Therefore,

$$G_{\Phi}^{\max}(\Delta)=\frac{a+c}{ac}=\frac{1}{a}+\frac{1}{c}.$$

Note that the optimal simulation uses three extreme points, $\Upsilon^{(1)}$, $\Upsilon^{(2)}$, and $\Upsilon^{(3)}$. It is not difficult to compute

$$G_{\Phi}^{\min}(\Delta) = \max\left\{\frac{1}{a} + \frac{1}{1-a}, \frac{1}{c} + \frac{1}{1-c}\right\}.$$

Since $a + c \leq 1$, $G_{\Phi}^{\max}(\Delta) \geq G_{\Phi}^{\min}(\Delta)$. ("=" holds if and only if $a + c = 1$.)

References

[1] S. Amari, Differential-geometrical methods in statistics, Lecture Notes in Statistics, 28 (1985).

[2] S. Amari and H. Nagaoka, Methods of Information Geometry,Translations of Mathematical Monograph,Vol.191(AM Sand Oxford University Press, 2000).

[3] A. Fujiwara, A Geometrical Study in Quantum Information Systems, doctoral thesis (1995).

[4] N. N. Cencov: Statistical Decision Rules and Optimal Inference. Trans. of Mathematical Monographs 53, Amer. Math. Soc., Providence (1982).

[5] K. Matsumoto, A Geometrical Approach to Quantum Estimation Theory, doctoral dissertation, University of Tokyo, 1998.

[6] D. Petz, Monotone metrics on matrix spaces, Linear Algebra Appl., 244,81–96 (1996).